Overview

Among the goals of this procurement is to minimize the execution time and maximize the overall throughput of multiple concurrent workstreams (see Large-Scale Computing in section C.4.1). Individual workstream components and some background materials are presented below. For the RFP response, the Offeror will use component measurements (where possible) and projections (where necessary) to propose delivered concurrent workstream throughput.

A Microsoft Excel spreadsheet has been supplied for reporting initial results. Report all performance measures using only the benchmark parallel framework. Report only actual measurements in the original table. Copy the table to report performance values obtained using alternate technologies; clearly mark projected values and provide details as to how the projection was derived.

Benchmark Components

The components described below comprise the software elements of the NOAA benchmark workstreams. Some optional standalone applications have been provided to assist the porting of the benchmark application to the target platform.

The 10 NOAA benchmark workstreams have been defined as:

- CM2 – Coupled Earth System Model (ESM)
- CM2 – Coupled High Resolution Model (HR)
- HIMF – Very High Resolution Ocean Model
- WRF-NMM 4.5 KM nonhydrostatic mesoscale model
- GFS T126 spectral global weather model
- GSI T254 Global Data Assimilation
- RUC-15km (ANX, pre-processing, forecast, and post processing)
- ROMS – oceanic model
- WRF 5KM atmospheric chemistry
- WRF 5KM static initialization.

Not all benchmark components are available at this time. See list below for availability. Some optional models have been included as an aid to porting. See individual comments below and additional instructions documents for workstream details.

Benchmark Model Overview

The CM2 – Earth System Model

This model is a core benchmark application. It is comprised of the N45L24 atmosphere core coupled to a 1-degree MOM4 ocean; land and ice model components are also run. While the atmosphere portion of the model is malleable with respect to layout and PE count, the best performance of our current production

model is achieved with a STATIC_MEMORY MOM4. Thus, a given executable may run multiple atmosphere configurations, but only one ocean layout. For example, the same executable may be used to run on 120 and 150 PEs in 60atm+60ocn or 90atm+60ocn concurrent mode; serial mode will always require a unique executable assuming that STATIC_MEMORY MOM4 shows performance advantages over the malleable form.

Multiple sample PE configurations have been provided. Concurrent mode examples carry the designation of the ocean portion of the model. Thus, cm2.30, cm2.60, cm2.90, cm2.120, cm2.150 and cm2.180 are all serial mode examples. Test cases cm2.30o.c (60PEs), cm2.60o.c (120 and 150PEs), cm2.72o.c (180PEs), cm2.80o.c (200PEs) and cm2.90o.c (180PEs) are all concurrent cases. All but cm2.72o.c and cm2.80o.c use executables also run for the serial cases; the atmosphere run on 72 or 80PEs is not interesting on current architectures and so these executables have not been run in serial mode.

One of the goals of a concurrent mode configuration should be load balance between the ocean model and remaining components. Until recently, the 60+60 and 90+90 configurations provided fairly good balance. Improvements in the time stepping scheme for MOM4 have just been introduced which change this balance. Moreover, it's expected the port to different architectures will produce different performance features for each of the model components. Thus, finding the best balance of processing element (PE) configurations will be part of the porting task.

Official RFI/RFP communication channels may be employed to request assistance with a new model decomposition configuration.

The CM2 – High Resolution Global Coupled Model

This model is a core benchmark application. It is comprised of the N90L40 atmosphere core (288 x 180 with 40 levels) coupled to a 1/3-degree MOM4 ocean (1080 x 840 with 75 levels); land and ice model components are also run. All comments from the CM2-ESM described above apply though of course the PE configurations are different. Owing to the much larger model sizes than the CM2-ESM case, there are far fewer tracers and diagnostics.

Official RFI/RFP communication channels may be employed to request assistance with a new model decomposition configuration.

As noted above, the ocean component of this coupled model is still in preparation at the time of initial benchmark set release. The standalone ocean and full coupled model will be made available as soon as possible.

Like the ESM version, the use of the –DSTATIC_MEMORY option requires each PE to have the same number of points in each of the horizontal directions.

The HIMF Very High Resolution Ocean Model

The drive to higher resolutions permeates climate research. The HIMF benchmark model is intended to be a first representative of the future classes of very high leading to ultra high resolution ocean models ($1/10^{th}$ degree and beyond).

The HIMF model is a core benchmark. It is a $1/6^{th}$ degree hemispheric model comprised of 2160 by 680 grid points. This model does not use the exchange grid and thus, bypasses one of the greatest present challenges to scalability. Thus, HIMF is shown to be highly scalable on current architectures and acts as a benchmark surrogate for the class of such codes.

The model is internally initialized, vastly reducing startup costs and input file size requirements. Even so, the benchmark consists of running but 2 simulation days. Lower resolution test cases are provided to aid porting. There is no requirement for the Offeror to run or report performance data for any of the lower resolution cases.

Like MOM4, the use of the –DSTATIC_MEMORY option requires each PE to have the same number of points in each of the horizontal directions.

As noted above, this model is still in preparation at the time of initial benchmark set release. It will be made available as soon as possible

Post Processing for workstreams 1, 2 and 3

The post processing data components are available as a smaller subset of the full benchmark and are supplied as an aid in porting to a target platform. The components are constructed to run on a single processing element (PE) using global history data model output in netCDF format. The components supplied as test cases are cpio/uncpio (pp1), splitvars (pp2), ncrcat (pp3), ncatted (pp4), ncks (pp5), timavg (pp6), ncap (pp7), plevel (pp8) and mppnccombine (pp9). These components contribute approximately 99% of the post processing wallclock time on the Origin machine.

Individual directories are provided for each component in the bench/run/pp directory. Each pp(n) directory contains data, scripts and output directories, and a run_pp(n).csh executable script. The data directory consists of representative data of varying file sizes. The scripts directory contains c-shell, awk and Bourne shell scripts for running each component and recording the average real, user and system time and total time for the input data of several file sizes. The output directory contains a pp(n)_times.txt file containing the times recorded on the current Origin 3000 platform. The output directory also contains the stdout file from tests. Many of the netCDF operators employed produce little testable output. Use the self tests which come with the netCDF and netCDF operator (NCO) libraries to confirm functionality.

To run each component, enter the run_pp(n).csh executable. After each run, an output text file, "out", is created in each pp(n) directory. This file contains the timing information for all the input file sizes. The average real, user and system times for each file size and the total time is written to stdout.

The directories containing the source code and makefiles or building the executables for pp2, pp6, pp8 and pp9 are in bench/build/pp.

The first component, pp1, contains timing information for packing and unpacking the cpio container file of global history data. For 100 years of post processing data for the CM2 model, cpio is executed 0 to 5 times for packing netCDF post processed data into a container file. Uncpio is executed between 3 and 22 times on the two 6-month global history files for each model component. The range of file sizes supplied for the test cases range from 24 to 384 MB. The actual file sizes for the full benchmark consist of the annual global history files and range from 24 MB to 5 GB. To aid in portability, the data is supplied as a gzipped tar file. The script will automatically unzip and untar the file.

The utility for extracting netCDF variables from data files into individual files is splitvars (pp2). It is executed between 1 and 3 times on static and concatenated annual global history data for 100 years of post processed CM2 data. The input data sizes provided are 33 to 528 MB. The full benchmark consists of static data files on the order of 100 KB and annual global history files between 24 MB and 5 GB.

The global history files are concatenated via the ncrcat NCO operator in the post processing scripts. This is represented in the third test case, pp3. ncrcat is executed between 5 and 1,826 times for 100 years of CM2 post processed data. The size distribution of the input data supplied ranges from 30 to 384 MB, while the full benchmark contains the annual global history files ranging from 24 MB to 5 GB.

The netCDF attribute editor, ncatted, is the fourth test case, pp4. The use of this utility is file size independent on the Origin 3000 test systems and it is executed between 1 and 1,498 times for 100 years of post processed CM2 data. The file sizes provided range from 30 MB to 1.9 GB for this component. The full benchmark sizes contain the annual global history files and range from 24 MB to 5 GB.

The fifth test case, pp5, represents the NCO operator, ncks. This is the kitchen sink utility for extracting subsets of netCDF files and it is executed on concatenated annual global history data between 12 and 2,490 times for 100 years of post processed CM2 data. The distribution of file sizes range from 30 MB to 1.9 GB. The full benchmark consists of annual global history files and range from 24 MB to 5 GB.

A component of the post processing stream, pp6, consists of time-averaging netCDF variables of concatenated global history files. For 100 years of post processed CM2 data, timavg is executed between 5 and 2,381 times. Time averaging is executed on varying file counts, depending on the frequency of the time average. Typical time averages occur on monthly, seasonal and annual scales. The file size supplied ranges from 33 MB to 528 MB, while the full benchmark input data ranges from 24 MB to 5 GB of annual global history data.

Arithmetic processing of netCDF files is carried out by the NCO operator ncap in test case pp7. It is executed only once for 100 years of post processed data in the initial script for the first year when the netCDF attributes are copied between monthly global history data files. The input file sizes provided range from 38 to 304 MB, while the full benchmark files are the monthly global history files and the sizes range from 2 to 430 MB.

The atmospheric data is processed on several atmospheric pressure levels in the test case pp8. For 100 years of post processed CM2 data, there are 17 pressure levels and plevel is executed between 5 and 19 times. The input data files are the concatenated global history files and range from 24 MB to 1.9 GB. The full benchmark consists of annual global history files and range from 24 MB to 5 GB.

Post processing benchmark 9 (pp9) is mppnccombine which is designed to concatenate individual process local domain history output into global domain history output.

The AM2P13 atmosphere core

The atmosphere is constructed to run on an "arbitrary" (though of course finite) number of processors using a single compilation. In general, the number of processing elements (PEs) and useful decompositions are limited by the number of grid points in each of the horizontal directions. There must be at least one grid point per PE for each of the model components (atmosphere, land, ice, ocean) or the model will fail at startup with an error. The Offeror is free to choose decompositions providing the best price performance for the model within the context of optimizing the workstream performance (see the SON for more information regarding "workstreams").

In the case of the standalone N45L24 and N90L40 models, a single executable will work for both models and all PE configurations. In general, configurations with equal numbers of points distributed to each PE are likely to perform best. The N45L24 atmosphere is 144 by 90 and the N45L40 model is 288 by 180. There is no requirement for the Offeror to run or report performance data for the standalone atmosphere cases.

It is important to note that optimal model decompositions for the standalone cases may not be optimal for the benchmark coupled model tests. The standalone cases favor decompositions with fewer PEs in the X direction because it reduces the overheads of the polar filter. When run in the fully coupled model, however, decompositions producing domains with roughly equal grid points in the X and Y (i.e. tending towards "square") are favored on our current platform due to the overheads of the exchange grid which couples the various model components.

The Offeror should also be aware that some routines, such as "data override", may be used in the standalone case where they are not used in the coupled model. Thus, the Offeror is cautioned to understand the profile of the fully coupled models prior to attempting optimizations.

The MOM4 ocean model

The ocean model has been shown to run substantially faster on a number of platforms using the STATIC_MEMORY preprocessing option. Fortran90 requires array components of derived types to be pointers. In general, compilers are rather conservative and what they assume about pointers and unnecessary temporary arrays tend to get generated. Fully specifying array extents in derived types removes the need for the pointer attribute at the cost of making the executable configuration fixed. Since the STATIC_MEMORY option improves MOM4 performance by a factor of almost 2x on it's current production platform, it is the mode in which it is generally run.

The resolution used for the CM2 – ESM is "1-degree" comprised of 360 by 200 grid points. An ocean standalone version of the model has been provided in the bench/mom4 directory. The standalone is provided to aid porting. There is no requirement for the Offeror to run or report performance data for the standalone ocean case.

The resolution used for the CM2 – HR is “1/3-degree” comprised of 1080 by 840 grid points. An ocean standalone version of the model will be provided in the bench/mom4 directory. The standalone is provided to aid porting. There is no requirement for the Offeror to run or report performance data for the standalone ocean case.

Note that the –DSTATIC_MEMORY option requires each PE to have the same number of points in each of the horizontal directions.

It is important to note that optimal model decompositions for the standalone cases may not be optimal for the benchmark coupled model tests.

The Offeror should also be aware that some routines, such as “data override”, may be used in the standalone case where they are not used in the coupled model. Thus, the Offeror is cautioned to understand the profile of the fully coupled models prior to attempting optimizations.

The WRF-NMM 4.5 KM nonhydrostatic mesoscale model

Not available at this time.

The GFS global weather model

This GFS model is a T126 spectral resolution with 64 levels in the vertical. GFS documentation can be found at

http://wwwt.emc.ncep.noaa.gov/gmb/moorthi/gam.html

The GFS has a malleable executable and runs on any number of PEs provided there is sufficient memory. The GFS is a hybrid MPI/OpenMP with the MPI task count and number of threads controlled at the script level. The GFS is reproducible across any number of PEs and varying numbers of MPI tasks and threads. An example using 10 nodes (10 MPI tasks with 3 threads) run on NCEP's IBM SP 1.3 Ghz Power 4 Cluster is provided. The GFS example script has a 48 hours forecast length and is controlled by the namelist variable FHSEG.

Instruction for building and running the GFS are contained in the README.126.64 file in the PORT directory of the tarfile.

GSI T254 Global Data Assimilation

Not available at this time.

RUC-15km (ANX, pre-processing, forecast, and post processing)

Not available at this time.

The ROMS – oceanic model

See http://quercus.igpp.ucla.edu/research/projects/roms/

The Regional Oceanic Modeling System (ROMS) is intended to be a coupled, multi-purpose, multi-disciplinary oceanic modeling tool.

There are two parts to the tests. The first part is a regression test verifying the same solution is independent of processor layout.

The second part is a long, benchmark run. Results should be reasonably close to those produced on the NOAA IJET Linux Cluster.

Not available at this time.

WRF 5KM with atmospheric chemistry

Not available at this time.

WRF 5KM with static initialization

This is a test of the Weather Research and Forecast (WRF) Advanced Research version (ARW).
The test contains six individual WRF tests with sample output and results for each. These six tests are: squall2d_x, squall2d_y, 3D quarter-circle shear supercell simulation, 2D flow over a bell-shaped hill, 3D baroclinic wave, and 2D gravity current. Each of these test simulations is described in the README_test_cases file, with an explanation of expected results.

Not available at this time.

Summary:

The fundamental build/run design is as follows:

0) Build and the base libraries necessary for each benchmark code (see individual benchmark documentation for library requirements).

1) Download bench.base.tgz. Unpack the benchmark build directory bench using gunzip and tar –xvf.

2) Change directory to bench and read the documentation specific to the workstream of interest.

NOTE: Default size of real must be 64; default size of integer may be either 32 or 64.

3) Follow workstream specific instructions to the benchmark build directories.

4) Edit the Makefile for local environment.

5) Make the executable; record the time make required to compile and link the executable in the spreadsheet provided.

6) Owing to their size, the benchmark run directories will be provided by tape. Offerors will supply a set of DLT format tapes to receive a copy of the benchmark run directories. Details for this process are pending and will be announced shortly.

There are 11 run directory components associated with the 10 benchmark workstreams. The pp post processing benchmark will serve as a surrogate for the post processing of the CM2-ESM, CM2-HR and HIMF workstreams.

Note there are some optional components intended only as an aid for porting.

- am2p13.esm.tgz	(optional)
- am2p13.hr.tgz	(optional)
- cm2.ems.tgz	(required)
- cm2.hr.tgz	(required - not available with first release)
- himf.vhr.tgz	(required - not available with first release)
- mom4.esm.tgz	(optional)
- mom4.hr.tgz	(optional - not available with first release)
- pp<1-9>.tgz	(post processing for workstreams 1, 2, and 3 - required)
- WRF-NMM_4.5_KM.tgz	(required - not available with first release)
- GFS.tgz	(required)
- GSI.tgz	(required - not available with first release)
- RUC.tgz	(required - not available with first release)
- ROMS.tgz	(required)
- WRFchem_5KM.tgz	(required - not available with first release)
- WRF+SI_5KM.tgz	(required)

Unpack the benchmark model run directory using gunzip and tar –xvf.

7) Copy the executable to the run directory.

8) Run the executable using appropriate MPI invocation; pipe stdout to a text file. For example, a 30PE run on our Origin system would look like:

mpirun –np 30 fms_cm2.30.x |& tee out.30

Of course any other means of capturing stdout is also acceptable. This copy of stdout other run output specified by specific directions should be returned with the benchmark data.

10) Report runtime as described for individual benchmark.

11) Report per process memory used.

12) Check results against verification files provided.

Model output files often have the same name regardless of PE count. Care must be taken not to accidentally overwrite output you wish to keep between consecutive model runs.

Note: The RFP benchmark will require a reproducibility test across PE count for some setting of the compiler, preferably that used for the performance benchmark itself. It is theoretically possible that high levels of compiler optimization might destroy reproducibility across PE count. To date, we have not seen this to be the case. Any mechanism causing loss of the capability to reproduce must be explained in detail. Use of such compiler optimizations will be weighed against the performance gain.

General Notes

This section is designed to give an overview of the items supplied with the benchmark. Details for compilation and model runs will be supplied with the benchmark directories.

Building the executables

The benchmark build directory structure has the following form:

bench/bin
bench/etc
bench/src.workstreams123
bench/src.workstreams456
bench/src.workstreams78910 (not available with first release)
bench/<model>

where <model> is: am2p13, cm2, gfs, mom4, roms, wrf_5km. The am2p13 and mom4 are standalone models provided to aid porting; there is no requirement for the vendor to run these cases.

Each <model> directory contains the exec/ and src/ subdirectories. Note that src/ is simply a soft link to bench/src.workstreams<nums>. Thus care must be exercised since modification of src/ for one model may affect code used by another. See workstream specific instructions for build specifications.

See workstream specific instructions for verification output is location.

Makefiles have been provided for each benchmark application. The Offeror is free to modify the Makefile directly or rebuild the Makefile to suit local conditions using tools provided with this benchmark.

Some models such as the am2p13 atmosphere can run with completely malleable executables (i.e. a single executable serves all PE configurations). Other codes such as the ocean models find significant performance improvements on some platforms when array components are fully specified at compile time through the STATIC_MEMORY preprocessing directive. See individual workstream instructions for application specifics.

Code optimizations

For the purposes of the benchmark, the following classes of “changes” for code optimization will be defined:

Class A: Modifications required to make a model run correctly, consistent with ANSI standard FORTRAN90 and C

Class B: Modifications to the program parallel communication

Class C: Modifications consistent with ANSI standard FORTRAN90/95 and C

Class D: All other modifications

Class A modifications are those required to allow a benchmark to run to completion correctly if, without such changes to source code, the benchmark will "fail" either by exiting prior to completion or producing incorrect answers. Class A modifications do not include any changes to source solely for performance.

Since there may be many causes for such changes (e.g. non-standard language usage within the application, workarounds required for compiler bugs, etc), the Government cannot state categorically that such modifications will not be evaluated without some sort of risk factor assigned. Still, it is the Government's desire to consider such changes as "essentially unmodified" code with no negative impact on evaluation.

Among the types of "changes" which will be taken as Class A are:

- Use of commercially supported libraries bid as part of the offering requiring no changes to benchmark source code or introduction of wrapper subroutines

- Compiler command lines with performance specific options including, but not limited to, automatic parallelization

- Automatic parallelization and multitasking mediated through the operating system

- Use of commercially available and supported source pre-processors bid as part of the offering.

Class B modifications are source code changes to the parallel communication libraries. These include use of communication libraries other than the benchmark provided parallel infrastructure.

Class C modifications are limited to those which do not reduce code portability and which remain consistent with ANSI standard FORTRAN90/95 and C (it is acknowledged that the codes as they exist may already contain some ANSI non-compliant features). Performance is important and the Government

is interested in performance enhancing code modifications. However, resources to implement and maintain such changes are limited. Thus while a risk assessment will be made of any such changes, they are encouraged.

Among the types of changes taken to be Class C are:

- Use of commercially supported libraries bid as part of the offering

- Use of compiler "directives" within the source

Class D modifications are all those changes to application source not included in Classes A, B, or C. Such modifications reduce code portability and tend to make development and maintenance more difficult and costly. **Class D modifications are very strongly discouraged**.

All acceptable changes must produce output consistent with the verification provided as described with each benchmark.

As described in the instructions below, baseline performance numbers comprised of only Class A modifications will be required. The benchmark supplied “parallel framework” will be required for this baseline where a communication library is employed (see the detailed instructions for definition of “parallel framework” with respect to a given benchmark). MPI (or any other communication library) is clearly not applicable for systems which use compiler or operating system mediated AUTOMATIC parallelization for the baseline benchmark. See workstream instructions for application specific details.

Offerors wishing to make code changes for evaluation must submit complete performance numbers for the workstream suite components containing the code changes IN ADDITION to the baseline numbers. Having satisfied the baseline requirement, the Offeror is free to mix classes of changes. Offerors are cautioned, however, that a set of performance numbers and the associated changes will be evaluated as a single entity and accepted or rejected as such.

While it is desirable, it is not required that the Offeror reach minimum performance requirements based on Class A changes alone. However, Offerors are again cautioned that source changes associated with a set of performance numbers are assessed risk as a single entity.

Running the model

Given the size of some of the input datasets, separate run directories have been provided as .tgz files (tar’d and gzipped). The Offeror will need to copy the executables built in the bench/ directories to the appropriate run directory. The run directories are named as the model targeted for that directory.

For the RFP response, the benchmark will require a reproducibility test across PE count for some setting of the compiler, preferably that used for the performance benchmark itself. It is theoretically possible that high levels of compiler optimization might destroy reproducibility across PE count. To date, we have not seen this to be the case. Any mechanism causing loss of the capability to reproduce must be explained in detail. Use of such compiler optimizations will be weighed against the performance gain.

Model output files often have the same name regardless of PE count. Care must be taken not to accidentally overwrite output you wish to keep between consecutive model runs.

See workstream instructions for benchmark specific details.

Reporting initial results

A reporting template in the form of a Microsoft Excel spreadsheet is provided. In general, the Offeror will report the total time required for the Makefile to run from a clean directory using the UNIX time command. Further, it is highly desirable for the Offeror to report measurements from 6 or more PE counts for each of the parallel benchmarks. As of the release date, only the CM2-ESM and atmosphere portion of CM2-HighRes, post processing for workstreams 1,2 and 3 and GFS are ready; the remaining portions will be released as quickly as possible. The PE counts should range from the lowest number of PEs on which the benchmark model can be run in a “reasonable period” (see individual benchmark instructions for more precise definitions) to the PE count past which the scaling curve is substantially flat.

Report value rounded to the nearest second.

At the Offeror’s option, a multiple additional template copies may be used to report a projection of the initial timings to target platforms the Offeror may propose.